Estimation of Fst and tree inference under hierarchical population structure

Tristan Mary-Huard*†[1] and David Balding[2]

[1]Institut national de la recherche agronomique [Paris-siège] (INRA Paris) – Institut national de la recherche agronomique [Paris-siège] – 147 rue de lÚniversité 75338 Paris Cedex 07, France
[2]University of Melbourne – Parkville VIC 3010, Australia

Abstract

Fst coefficients measures the genetic differentiation among a set of populations. There has been confusion/disagreement for decades over the definition of the Fst coefficients that can be based either on correlations of pairs of alleles sampled across populations (Weir and Hill definition, noted FstWH hereafter) or on mismatch probabilities within and between the sampled populations (Hudson definition, noted FstH hereafter). On simple models, the two definitions have been shown to be consistent.

Here we consider a hierarchical population structure model, assuming a hierarchical set of ancestral populations represented in a tree. In this setting explicit expressions of both the FstWH and the FstH coefficients can be obtained in terms of the tree branch lengths. These expressions highlight the fact that the two Fst definitions are generally not consistent, and capture complementary properties of the evolutionary history of the populations.

We then present an efficient procedure to jointly infer the tree structure and its associated coefficients. First at a given locus, the simple moment estimator $f_k(1-f_{k'})$ can be computed, where f_k and $f_{k'}$ are the empirical allelic frequencies in populations k and k', respectively. This moment estimator is shown to capture the information about the tree branch lengths involved in the common history of the 2 populations. Assuming a shared history, moment estimators may be averaged over all loci. This provides us with an averaged estimate for all pairs of population that can be used to infer the tree as follows: at each step an intermediate ancestral population is added between the ancestral population common to all observed populations and one of its child populations. The algorithm stops when a binary tree is obtained. At the jth step, such a hierarchical clustering strategy requires the solving of $K-j
(2constrainedoptimizations,sothatintotal${O}(K3)$optimizationsareperformed.Sincealloftheseoptimizationsarebasedonthesameaveragemom

*Speaker
†Corresponding author: maryhuar@agroparistech.fr